Review of maximum likelihood estimation

560 Hierarchical modeling

Peter Hoff

Statistics, University of Washington
lmer(formula, data = NULL, REML = TRUE, 
    control = lmerControl(), start = NULL, verbose = 0L, 
    subset, weights, na.action, offset, contrasts = NULL, 
    devFunOnly = FALSE, ...)
```r
library(lme4)

lmer(y~1+(1|g))
```

```r
## Linear mixed model fit by REML ['lmerMod']
## Formula: y ~ 1 + (1 | g)
## REML criterion at convergence: 177.9876
## Random effects:
## Groups   Name            Std.Dev.
## g        (Intercept)    0.6197
## Residual                 1.3369
## Number of obs: 50, groups: g, 10
## Fixed Effects:
## (Intercept)
##     16.31
```
Method of moments

```r
aovfit <- aov(lm(y ~ as.factor(g)))
MSG <- aovfit[1, 3]
MSE <- aovfit[2, 3]
t2 <- (MSG - MSE) / n
s2 <- MSE
t2
## [1] 0.3840768
s2
## [1] 1.787206
sqrt(t2)
## [1] 0.6197393
sqrt(s2)
## [1] 1.336864
mean(y)
## [1] 16.30641
```
A more complicated example

```r
nels_mathdat[1:10,]
```

<table>
<thead>
<tr>
<th></th>
<th>school</th>
<th>enroll</th>
<th>flp</th>
<th>public</th>
<th>urbanicity</th>
<th>hwh</th>
<th>ses</th>
<th>mscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1011</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>urban</td>
<td>2</td>
<td>-0.23</td>
<td>52.11</td>
</tr>
<tr>
<td>2</td>
<td>1011</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>urban</td>
<td>0</td>
<td>0.69</td>
<td>57.65</td>
</tr>
<tr>
<td>3</td>
<td>1011</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>urban</td>
<td>4</td>
<td>-0.68</td>
<td>66.44</td>
</tr>
<tr>
<td>4</td>
<td>1011</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>urban</td>
<td>5</td>
<td>-0.89</td>
<td>44.68</td>
</tr>
<tr>
<td>5</td>
<td>1011</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>urban</td>
<td>3</td>
<td>-1.28</td>
<td>40.57</td>
</tr>
<tr>
<td>6</td>
<td>1011</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>urban</td>
<td>5</td>
<td>-0.93</td>
<td>35.04</td>
</tr>
<tr>
<td>7</td>
<td>1011</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>urban</td>
<td>1</td>
<td>0.36</td>
<td>50.71</td>
</tr>
<tr>
<td>8</td>
<td>1011</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>urban</td>
<td>4</td>
<td>-0.24</td>
<td>66.17</td>
</tr>
<tr>
<td>9</td>
<td>1011</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>urban</td>
<td>8</td>
<td>-1.07</td>
<td>46.17</td>
</tr>
<tr>
<td>10</td>
<td>1011</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>urban</td>
<td>2</td>
<td>-0.10</td>
<td>58.76</td>
</tr>
</tbody>
</table>
A more complicated example

\[ y_{i,j} = (\beta_0 + \beta_{0,j}) + \beta_1 \times flp_j + \beta_2 \times enroll_j + (\beta_3 + \beta_{3,j}) \times ses_{i,j} + \epsilon_{i,j} \]

```r
fit<-lmer(mscore~flp+enroll+ses+(ses|school),data=nels_mathdat,REML=FALSE)
```
summary(fit)

## Linear mixed model fit by maximum likelihood ['lmerMod']
## Formula: mscore ~ flp + enroll + ses + (ses | school)
## Data: nels_mathdat
##
## AIC   BIC   logLik  deviance  df.resid
## 92397.7 92457.5 -46190.9 92381.7       12966
##
## Scaled residuals:
## Min 1Q Median 3Q Max
## -3.9797 -0.6399 0.0180 0.6681 4.5053
##
## Random effects:
## Groups   Name    Variance  Std.Dev.  Corr
## school   (Intercept)  9.004   3.001
##           ses         1.600   1.265   0.05
## Residual             67.260   8.201
## Number of obs: 12974, groups: school, 684
##
## Fixed effects:
## Estimate  Std. Error t value
## (Intercept) 55.429339   0.402907 137.57
## flp         -2.411519   0.185311  -13.01
## enroll      0.007095   0.082023   0.09
## ses         4.116886   0.125381   32.83
##
## Correlation of Fixed Effects:
##          (Intr) flp  enroll
## flp   -0.815
## enroll -0.300 -0.193
## ses    -0.202  0.212  0.007
Models and inference

A statistical model is a collection of probability distributions for observed data:

\[ \mathcal{P} = \{ p(y|\theta), \theta \in \Theta \} \]

- \( y \) is the data;
- \( \Theta \) is the set of parameter values;
- \( p(y|\theta) \) is a probability (density) for each \( \theta \in \Theta \).
Example: Normal model

For example, the normal model is

\[ p(y|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{(y - \mu)^2}{2\sigma^2} \right\}, \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+ \].

- \( y \) is a single observed data value;
- \( \theta = \{\mu, \sigma^2\} \) is the parameter (or are the parameters);
- \( \Theta = \mathbb{R} \times \mathbb{R}^+ \) is the set of possible parameter values;
- \( p(y|\mu, \sigma^2) \) is the normal probability density for each \( \mu, \sigma^2 \).
Example: Normal model
Model-based inference

*Model-based statistical inference* involves

**Estimation**: Obtaining a value \( \hat{\theta} \in \Theta \) that “best” represents the population.

**Inference**: Describing how well \( \hat{\theta} \) represents the population.

Inference includes things like: confidence intervals, hypotheses tests.

*Likelihood-based statistical inference*:

- a type of model based inference;
- estimation and inference are based on the likelihood function.
Joint probability of the data

**Independent events:** Recall if $A$ and $B$ are independent events,

$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B).$$

**Independent observations:** If $y_1$ and $y_2$ are independent observations, then

$$p_{y_1y_2}(y_1, y_2 | \theta) = p(y_1 | \theta) \times p(y_2 | \theta)$$

$$= \prod_{i=1}^{2} p(y_i | \theta)$$

**Independent sample:** If $y = (y_1, \ldots, y_n)$ are independent observations, then

$$p_y(y | \theta) = p(y_1 | \theta) \times \cdots \times p(y_n | \theta)$$

$$= \prod_{i=1}^{n} p(y_i | \theta)$$

$p_y(y | \theta)$ is the *joint probability (density)* of the data.
Example: Binary data

Suppose we are sampling people from a population and recording whether or not they have a particular disease.

Let \( y_i \in \{0, 1\} \) depending on if they are uninfected or infected.

A natural model is the binomial/binary model:

\[
y_1, \ldots, y_n \sim \text{i.i.d. binary}(\theta), \quad \theta \in [0, 1]
\]

In this model

- The parameter is \( \theta \in [0, 1] \).
- The probability density is

\[
p(y|\theta) = \begin{cases} 
(1 - \theta) & \text{if } y = 0 \\
\theta & \text{if } y = 1 
\end{cases},
\]

which can be compactly written as

\[
p(y|\theta) = \theta^y (1 - \theta)^{1-y}.
\]
Joint probability

If \( y_1, \ldots, y_n \) are i.i.d. samples from this population,

\[
p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta)
\]

\[
= \prod_{i=1}^{n} \theta^{y_i} (1 - \theta)^{1-y_i}
\]

\[
= \theta^{\sum y_i} (1 - \theta)^{n - \sum y_i}
\]

**Interpretation:**
\( p(y|\theta) \) tells you how probable a given outcome is, for a particular \( \theta \).
Binary sequence probabilities

Quiz: If $n = 3$ and $\theta = 1/2$, what is

- $p(\{1, 0, 1\}|\theta)$?
- $p(\{0, 0, 0\}|\theta)$?

Quiz: If $n = 3$ and $\theta = 1/3$, what is

- $p(\{1, 0, 1\}|\theta)$?
- $p(\{0, 0, 0\}|\theta)$?

Foreshadowing:
If your observed data were $\{0, 0, 0\}$, which $\theta$ value is “more likely”? 
The **likelihood** is the probability of the data as a function of the parameter:

\[ L(\theta : y) = p(y|\theta) \]

**Example (binomial model):** If \( y = \{0, 0, 0\} \), then

\[ L\left(\frac{1}{2} : \{0, 0, 0\}\right) = \frac{1}{8} = 0.125 \]

\[ L\left(\frac{1}{3} : \{0, 0, 0\}\right) = \frac{8}{27} \approx 0.296 \]

We say \( \{\theta = 1/3\} \) has a higher likelihood than \( \{\theta = 1/2\} \) for these data.
The **maximum likelihood estimator**, or **MLE**, is the value of \( \theta \) that maximizes the likelihood:

\[
\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\text{arg max}} \ L(\theta : y)
\]

**Example (binomial model):** If \( y = \{0, 0, 0\} \) and \( \theta \) is either \( 1/2 \) or \( 1/3 \), then

\[
\Theta = \{1/3, 1/2\}
\]

\[
\hat{\theta}_{MLE} = 1/3
\]

because \( L(1/3 : \{0, 0, 0\}) > L(1/2 : \{0, 0, 0\}) \).
Binomial MLE

Suppose 5 people are infected in a sample of size 30.

\[ n = 30, \quad \sum y_i = 5 \]

The likelihood function is

\[ L(\theta : y) = \theta^{\sum y_i} (1 - \theta)^{n - \sum y_i} = \theta^5 (1 - \theta)^{25}. \]

Careful examination, or trial and error gives \( \hat{\theta} = 5/30 = 1/6 = 0.166\overline{6}. \)
Binomial MLE

Suppose 50 people are infected in a sample of size 300.

\[ n = 300, \sum y_i = 50 \]

The likelihood function is

\[ L(\theta : y) = \theta^{\sum y_i} (1 - \theta)^{n - \sum y_i} = \theta^{50} (1 - \theta)^{250}. \]

Careful examination, or trial and error gives \( \hat{\theta} = 50/300 = 1/6. \)
Log likelihoods

Likelihoods with lots of data can give extreme numbers.

Alternatively, we can make inference with the **log-likelihood**:

If $\hat{\theta}$ maximizes $L(\theta : y)$ then it also maximizes $\log L(\theta : y) = l(\theta : y)$.

To find the MLE we can work with the log-likelihood. For the binomial model,

$$L(\theta : y) = \theta^{\sum y_i} (1 - \theta)^{n - \sum y_i}$$

$$l(\theta : y) = \log \left( \theta^{\sum y_i} (1 - \theta)^{n - \sum y_i} \right)$$

$$= \log \theta^{\sum y_i} + \log (1 - \theta)^{n - \sum y_i}$$

$$= (\sum y_i) \times \log \theta + (n - \sum y_i) \times \log (1 - \theta)$$
Binomial MLE

Suppose 5 people are infected in a sample of size 30.

\[ n = 30, \sum y_i = 5 \]

The log-likelihood function is

\[
 l(\theta : y) = 5 \times \log(\theta) + 25 \times \log(1 - \theta).
\]

As before, \( \hat{\theta} = \frac{5}{30} = \frac{1}{6} = 0.166\bar{6} \).
Binomial MLE

Suppose 50 people are infected in a sample of size 300.

\[ n = 300, \sum y_i = 50 \]

The log-likelihood function is

\[ l(\theta : y) = 50 \times \log(\theta) + 250 \times \log(1 - \theta). \]

As before, \( \hat{\theta} = \frac{5}{30} = \frac{1}{6} = 0.166\bar{6}. \)
Comparing log-likelihoods
Inference with the likelihood function

As we’ve seen and discussed,
- the peak of the log-likelihood gives the MLE.
- the curvature of the log-likelihood gives the *information* or *certainty*.

How can we find the peak in general?

What is the information? How does it relate to estimation accuracy?
Recall from calculus that the *tangent* or *derivative* of a function, at a local maximum, will be zero. This tells us how to find the MLE:

\[
\hat{\theta}_{MLE} \text{ satisfies } \frac{d}{d\theta} l(\theta : y) |_{\theta = \hat{\theta}} = 0
\]

Let’s try this for the binomial model. Recall that

\[
\frac{d}{d\theta} \log \theta = \frac{1}{\theta}, \quad \frac{d}{d\theta} \log(1 - \theta) = -\frac{1}{1 - \theta}
\]

The derivative of the log-likelihood is

\[
\frac{d}{d\theta} l(\theta : y) = \frac{d}{d\theta} \left( \sum y_i \times \log \theta + (n - \sum y_i) \times \log(1 - \theta) \right)
\]

\[
= \sum \frac{y_i}{\theta} - \frac{n - \sum y_i}{1 - \theta}
\]
Finding the MLE

Therefore

\[ \frac{d l(\theta : y)}{d \theta} \bigg|_{\theta = \hat{\theta}} = \frac{\sum y_i}{\hat{\theta}} - \frac{n - \sum y_i}{1 - \hat{\theta}} = 0 \text{ if} \]

\[ \frac{\sum y_i}{\hat{\theta}} = \frac{n - \sum y_i}{1 - \hat{\theta}} \]

\[ \sum y_i - \hat{\theta} \sum y_i = \hat{\theta} n - \hat{\theta} \sum y_i \]

\[ \hat{\theta} = \frac{\sum y_i}{n} \]

So not surprisingly, the MLE is the sample proportion \( \frac{\sum y_i}{n} \).
Information and precision

The precision of the MLE (how well it estimates the truth) depends on the second derivative, or curvature, of the log-likelihood.

For the binomial model, the second derivative is

$$\frac{d^2 l(\theta : y)}{d \theta^2} = -\frac{\sum y_i}{\theta^2} - \frac{n - \sum y_i}{(1 - \theta)^2}$$

Plugging in the MLE $\hat{\theta}$ for $\theta$ gives

$$\frac{d^2 l(\theta : y)}{d \theta^2} \bigg|_{\theta = \hat{\theta}} = -\frac{n}{\hat{\theta}} - \frac{n}{(1 - \hat{\theta})} = -\frac{n}{\hat{\theta}(1 - \hat{\theta})}$$

Information: In stat theory, the observed information about $\theta$ is

$$I_n = -\frac{d^2}{d \theta^2} l(\theta : y) \bigg|_{\hat{\theta}} = \frac{n}{\hat{\theta}(1 - \hat{\theta})}$$

for the binomial model

Exercise: Consider how $I_n$ varies with $n$ and $\hat{\theta}$. 
Information, variance and CIs

In many problems, the inverse of the information gives a variance estimate:

\[
\text{Var}[\hat{\theta}] \approx \frac{1}{I_n}
\]
\[
\text{sd}(\hat{\theta}) \approx \sqrt{\frac{1}{I_n}}
\]
\[
\text{se}(\hat{\theta}) = \sqrt{\frac{1}{I_n}}
\]

For the binomial model, \( I_n = \frac{n}{[\hat{\theta}(1 - \hat{\theta})]} \), so

\[
\text{Var}[\hat{\theta}] \approx \frac{\hat{\theta}(1 - \hat{\theta})}{n}
\]
\[
\text{sd}(\hat{\theta}) \approx \sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}}
\]
\[
\text{se}(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}}
\]

An approximate 95% CI for \( \theta \) is then

\[
\hat{\theta} \pm 2\sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}}.
\]

This is known as the “Wald interval” for a binomial proportion.
MLE for the hierarchical normal model

\[ y_{i,j} = \mu + a_j + \epsilon_{i,j} \]
\[ \{\epsilon_{i,j}\} \sim \text{iid } \mathcal{N}(0, \sigma^2) \]
\[ \{a_j\} \sim \text{iid } \mathcal{N}(0, \tau^2) \]

Parameters to estimate:

- Fixed effects: \( \mu \)
- Variance components: \( \sigma^2, \tau^2 \)
- Random effects: \( a_1, \ldots, a_m \)

Likelihood estimation focuses on estimation of \( \theta = (\mu, \sigma^2, \tau^2) \)

Alternative methods are required for estimation of \( a_1, \ldots, a_m \).
HNM likelihood

Data:

\[ \mathbf{y} = (y_{1,1}, \ldots, y_{nj,1}, \ldots, y_{1,m}, \ldots, y_{nm,m}) \]
\[ = (\{y_{1,1}, \ldots, y_{nj,1}\}, \ldots, \{y_{1,m}, \ldots, y_{nm,m}\}) \]
\[ = (y_1, \ldots, y_n) \]

Likelihood:

\[ l(\mu, \sigma^2, \tau^2 : \mathbf{y}) = p(\mathbf{y} | \mu, \tau^2, \sigma^2) \]

Recall: Under the HNM,

- observations within groups are correlated;
- observations across groups are independent.

\[ l(\mu, \sigma^2, \tau^2 : \mathbf{y}) = p(\mathbf{y} | \mu, \tau^2, \sigma^2) = p(y_1 | \mu, \tau^2, \sigma^2) \times \cdots \times p(y_m | \mu, \tau^2, \sigma^2) \]
\[ = \prod_{j=1}^{m} p(y_j | \mu, \tau^2, \sigma^2) \]
Likelihood contribution from a single group

\[ y_{i,j} = \mu + a_j + \epsilon_{i,j} \]

\[ \epsilon_{1,j}, \ldots, \epsilon_{n_j,j} \sim \text{iid } N(0, \sigma^2) \]

\[ a_j \sim N(0, \tau^2) \]

As we've discussed, the \( y_{i,j} \)'s are normal with

- \( \mathbb{E}[y_{i,j} | \mu] = \mu \)
- \( \text{Var}[y_{i,j} | \mu] = \sigma^2 + \tau^2 \)
- \( \text{Cov}[y_{i_1,j}, y_{i_2,j}] = \tau^2 \)

In vector form, we can express this as follows:

\[
\mathbb{E}[\mathbf{y}_j | \mu] = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} = \mu \mathbf{1} \\
\text{Cov}[\mathbf{y}_j | \mu] = \begin{pmatrix} \sigma^2 + \tau^2 & \tau^2 & \cdots & \tau^2 \\ \tau^2 & \sigma^2 + \tau^2 & \cdots & \tau^2 \\ \vdots & \vdots & \ddots & \vdots \\ \tau^2 & \tau^2 & \cdots & \sigma^2 + \tau^2 \end{pmatrix}
\]
Multivariate normal distribution

This means that $y_j$ has a multivariate normal distribution.

The density of a general multivariate normal($\mu, \Sigma$) distribution is

$$p(y|\theta, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\{-{(y - \theta)^T\Sigma^{-1}(y - \theta)/2}\}$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma^2_1 & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{1,2} & \sigma^2_2 & \cdots & \sigma_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,p} & \sigma_{2,p} & \cdots & \sigma^2_p \end{pmatrix}.$$
Computing the log-likelihood

MLEs of \((\mu, \sigma^2, \tau^2)\) can be found by maximizing the log likelihood.

**Log likelihood:**

\[
L(y : \mu, \sigma^2, \tau^2) = p(y_1, \ldots, y_m | \mu, \sigma^2, \tau^2)
\]

\[
l(y : \mu, \sigma^2, \tau^2) = \log p(y_1, \ldots, y_m | \mu, \sigma^2, \tau^2)
\]

\[
= \log \prod_{j=1}^{m} p(y_j | \mu, \sigma^2, \tau^2)
\]

\[
= \sum_{j=1}^{m} \log p(y_j | \mu, \sigma^2, \tau^2),
\]

where \(\log p(y_j | \mu, \sigma^2, \tau^2)\) is the log of a multivariate normal density.

For the HNM, we replace

- \(\theta\) with \(\mu 1\)
- \(\Sigma\) with the covariance matrix from the previous slide.
Computing the (minus) log-likelihood

mll.oneway

## function(mus2t2,y,g)
## {
## ll<-0
## for(gj in sort(unique(g)))
## {
## nj<-sum(g==gj)
## S<-diag(s2,nj) + matrix(t2,nj,nj)
## ll<-ll+ldmvnorm(y[g==gj],mu,S)
## }
## -ll
## }

Example: Wheat data

```r
mll.oneway( c(16.3, 1.787, 0.384 ), y,g )
## [1] 88.6121
mll.oneway( c(15, 1.787, 0.384 ), y,g )
## [1] 100.1217
mll.oneway( c(16.3, 2, 0.384 ), y,g )
## [1] 88.76881
mll.oneway( c(16.3, 1.787, 0.3 ), y,g )
## [1] 88.58599
mll.oneway( c(16.3, 1.787, 0.2 ), y,g )
## [1] 88.67161
```
The MLEs are

\[ \hat{\mu} = 16.3063995, \; \hat{\sigma}^2 = 1.7872063, \; \hat{\tau}^2 = 0.3099255 \]
Confidence intervals via the Information matrix

For maximum likelihood estimation in general,

- $\hat{\theta}_{MLE} \to \theta$ as the sample size goes to infinity (if the model is correct);
- $\hat{\theta} \sim \text{normal}(\theta, \text{Var}[\hat{\theta}])$, where
- $\text{Var}[\hat{\theta}] \approx -\left[\frac{d^2 l(\theta | y)}{d\theta^2}\right]^{-1}$ for large sample sizes.

For our hierarchical normal model, this means that approximate 95% confidence intervals for $(\mu, \tau^2, \sigma^2)$ can be obtained from the curvature of the log likelihood.
Confidence intervals via the Information matrix

The observed information matrix is the (matrix of) second derivative(s) of the negative log-likelihood function at the MLE (aka the Hessian):

\[
I_n(\hat{\theta} : y) = \left\{ -\frac{\partial^2 l(\theta : y)}{\partial \theta_j \partial \theta_k} \right\}_{\theta = \hat{\theta}}
\]

The inverse of the information matrix gives an estimate of the variance/covariance of the MLE’s:

\[
\text{Var}[\hat{\theta} : y] \approx I_n^{-1}(\hat{\theta} : y)
\]

From this, we can get confidence intervals:

- \( \sqrt{I_{jj}^{-1}} \) gives an approximate standard error for \( \theta_k \).
- The MLE plus and minus 2 standard errors gives a rough confidence interval for the parameters.

\[
\text{Pr}(\theta \in \hat{\theta} \pm 2 \times \text{se}[\hat{\theta}]) \approx 0.95
\]
Confidence intervals via the Information matrix

theta.wheat<-fit.ml$par

theta.wheat

## [1] 16.3063995 1.7872063 0.3099255

I<-fit.ml$hessian

V.wheat<-solve(I)

V.wheat

## [,1]       [,2]       [,3]
## [1,]  6.673668e-02 -5.694851e-11 -6.482475e-08
## [2,] -5.694851e-11  1.597051e-01  3.194081e-02
## [3,] -6.482475e-08 -3.194081e-02  9.546274e-02

sqrt(diag(V.wheat))

## [1] 0.2583344 0.3996312 0.3089705

theta.wheat+2*sqrt(diag(V.wheat))

## [1] 16.8230684 2.5864686 0.9278664

theta.wheat-2*sqrt(diag(V.wheat))

## [1] 15.7897307 0.9879440 -0.3080154
NELS example

100 randomly sampled schools from the NELS dataset
Analysis of all schools

```
fit.ml.nels<-optim(c(50, 1, 1), mll.oneway, gr = NULL, y = mscores, g = schools, lower = c(-Inf, 0, 0), method = "L-BFGS-B", hessian = TRUE)

fit.ml.nels

## $par
## [1] 50.93914 73.70881 23.63382
##
## $value
## [1] 46956.63
##
## $counts
## function gradient
##  27  27
##
## $convergence
## [1] 0
##
## $message
## [1] "CONVERGENCE: REL_REDUCTION_OF_F <= FACTR*EPSMCH"
##
## $hessian
## [,1]    [,2]    [,3]
## [1,] 24.35837 -0.01577  0.04914
## [2,] -0.01577  1.13128  0.03027
## [3,]  0.04914  0.03027  0.42090
```

The MLEs are

\[
\hat{\mu} = 50.9391407 \, , \, \hat{\sigma^2} = 73.708808 \, , \, \hat{\tau^2} = 23.6338229
\]
Confidence intervals via the Information matrix

```
theta.nels <- fit.ml.nels$par
theta.nels
## [1] 50.93914 73.70881 23.63382
I <- fit.ml.nels$hessian
V.nels <- solve(I)
V.nels
## [,1]       [,2]       [,3]
## [1,] 0.0410638760 0.0007019913 -0.004844505
## [2,] 0.0007019913 0.8856698641 -0.063767034
## [3,] -0.0048445047 -0.0637670344 2.381014344
sqrt(diag(V.nels))
## [1] 0.2026422 0.9411003 1.5430536
theta.nels + 2 * sqrt(diag(V.nels))
## [1] 51.34443 75.59101 26.71993
theta.nels - 2 * sqrt(diag(V.nels))
## [1] 50.53386 71.82661 20.54772
```
Fitting via `lme4`: Wheat

```r
fit.wheat <- lmer(yield ~ 1 + (1 | region), REML = FALSE)
summary(fit.wheat)

## Linear mixed model fit by maximum likelihood ['lmerMod']
## Formula: yield ~ 1 + (1 | region)
##
## AIC      BIC   logLik deviance df.resid
## 183.2 188.9  -88.6  177.2       47
##
## Scaled residuals:
##    Min  1Q Median  3Q   Max
## -2.7913 -0.6035  0.1311  0.6520  1.7262
##
## Random effects:
## Groups   Name        Variance Std.Dev.
## region   (Intercept) 0.3099    0.5567
## Residual             1.7872    1.3369
## Number of obs: 50, groups: region, 10
##
## Fixed effects:
##                Estimate Std. Error   t value
## (Intercept) 16.3064     0.2583    63.12

theta.wheat

## [1] 16.3063995 1.7872063 0.3099255

sqrt(diag(V.wheat))

## [1] 0.2583344 0.3996312 0.3089705
```
Fitting via lme4: Schools

```r
code
fit.nels <- lmer(mscores ~ 1 + (1 | schools), REML = FALSE)
summary(fit.nels)
```

```
## Linear mixed model fit by maximum likelihood ['lmerMod']
## Formula: mscores ~ 1 + (1 | schools)
##
## AIC   BIC  logLik deviance df.resid
## 93919.3 93941.7 -46956.6 93913.3    12971
##
## Scaled residuals:
##    Min 1Q Median 3Q Max
## -3.8112 -0.6534  0.0093  0.6732  4.6999
##
## Random effects:
## Groups   Name  Variance  Std.Dev.
## schools (Intercept)     23.63     4.861
## Residual                      73.71     8.585
## Number of obs: 12974, groups: schools, 684
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)      50.939     0.2026  251.4
```

```
theta.nels
```

```
## [1] 50.93914 73.70881 23.63382
```

```
sqrt(diag(V.nels))
```

```
## [1] 0.2026422 0.9411003 1.5430536
```
Our technology so far

**ANOVA, method of moments:**
- Estimation: $\hat{\mu} = \bar{y}_\cdot$, $\hat{\sigma}^2 = MSE$, $\hat{\tau}^2 = (MSG - MSE)/n$
- Inference: $F$-test for across-group differences.

**Maximum likelihood:**
- Estimation: MLEs $\hat{\mu}$, $\hat{\sigma}^2$, $\hat{\tau}^2$
- Inference: CIs via likelihood curvature.

What about estimation of $a_j$ or $\mu_j$'s?
Estimation of group level means

We will consider two types of estimates of the $\mu_j$’s:

**Unbiased sample mean estimates:**

$$\hat{\mu}_j = \bar{y}_j$$

**Biased shrinkage estimates:**

$$\hat{\mu}_j = \frac{n_j / \hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\tau}^2} \bar{y}_j + \frac{1 / \hat{\tau}^2}{\hat{\sigma}^2 + 1 / \hat{\tau}^2} \bar{y}_\cdot$$

The latter will be preferable when $\tau^2$ is small compared to $\sigma^2 / n_j$. 